RESAMPLING METHODS FOR MARKOV PROCESSES WITH NO MIXING CONSTRAINTS

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

ELECTRICAL ENGINEERING

DECEMBER 2017

By

Kevin Oshiro

Thesis Committee:

Narayana Santhanam, Chairperson

James Yee

June Zhang

Acknowledgements

First, I would like to thank my adviser Dr. Narayana Santhanam for his support and guidance. He has provided me with valuable lessons on many occasions and I am grateful to have had the opportunity to learn from him. I would also like to thank my committee members Dr. James Yee and Dr. June Zhang for their feedback on my research and for their advice in general. Furthermore, I would like to express my gratitude to fellow graduate students Maryam and Changlong for their occasional assistance and useful discussions regarding research and classes. I am also grateful to my friends for their encouragement and support in pursuing my academic goals. Finally, I would like to thank my parents for their love, support, and patience. It is because of them that I have been able to accomplish all that I have done.

Abstract

Jackknife and bootstrap are resampling procedures that can be used to reduce the bias or estimate the variance of a statistic. These methods are useful because they perform well and are simple to implement, but an important assumption for their good performance is that of i.i.d. sampling. Previous analysis of these techniques for processes with memory generally require constraints on the memory or the mixing.

In this work we adapt the jackknife and bootstrap procedures to estimate the variance of conditional probability estimates when we have unbounded memory and make no assumptions on the mixing of the process. We only require that the process satisfies a continuity condition, which says that the incremental value of a bit in the past diminishes with increasing distance. We then analyze the procedures to provide bounds on the bias of the estimates.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Suppose we have a finite number of samples drawn from an unknown distribution, and using these samples we estimate some parameter of that distribution. One important question to ask is whether the estimate is good or not. In other words, how close do we expect it to be to the true value and if we were to obtain another sample and estimate again, how much would we expect our answer to vary? Of course, we can't provide answers without knowing the true distribution, but what we can do is use the data to estimate these values. For this we can use resampling procedures such as jackknife or bootstrap. These techniques are favorable because they are easy to understand, straight foward to implement, and require few assumptions.

Jackknife allows one to reduce the bias or estimate the variance of a sample via recalculated statistics, obtained by sequentially omitting data points. Bootstrap is used to estimate the properties of an estimator by drawing samples from the empirical distribution given by the sample. Cross-validation also falls into the category of resampling procedures, but it is used to estimate of the error associated with a predictive model built on data by systematically rotating different subsets of the data between training and validation.

One assumption which is essential for obtaining good results is that of independent sampling. It is easy to see that ignoring dependencies between the samples can give unfavorable results. Much work has been done to adapt these resampling procedures to general Markov and stationary ergodic setups. One group of results, [6], [7], [8], [9], [10], [11], and [12], considers grouping the samples into blocks. The jackknife procedures then operates by removing the blocks in some systematic manner,

while the bootstrap samples from the set of blocks. Another line of research, [13], [14], [15], [16], considers estimating the transition probabilities of the source and then generating new samples using the estimated source. However, these approaches generally require some sort of assumption on the mixing of the process, i.e., how quickly the empirical frequencies in the sample reflect the stationary probabilities of the source.

In this work we make no assumptions on the mixing of the process or the length of the memory. With no assumptions the problem is ill-posed, as Markov souces can have long enough memory and sufficiently slow mixing such that they are indistinguishable from an i.i.d. source for any sample size. Therefore we adopt a continuity condition, as in [1], which says that bits that are farther apart, conditioned on all bits between, have less information about each other. In other words, our condition acts as a soft memory constraint and does not at all constrain the mixing.

Given this setup, we apply the jackknife procedure to estimate the variance of the empirical transition probabilities and give a bound on the bias of the estimate. We also apply the bootstrap procedure for the same estimates and give a bound on the bias of the bootstrap estimate. Additionally, we will see that some interesting phenomena arise in the slow mixing regime; e.g., estimates for longer contexts can sometimes be better than estimates of their suffixes, despite having less data.

Motivation

To motivate the problem of estimating parameters of slow mixing Markov processes, we consider the following example. Suppose we have a hypothetical user browsing the internet for political news. Starting at one news site, the user follows one of the available hyperlinks to another site, then follows a link from that site, and so on. As many political news sites tend to be biased more liberal or conservative, they likely link mostly to other pages with a similar persuasion. Additionally, the beliefs held by the user influence which links are chosen. So the user will likely be confined to a certain subset of webpages and will obtain news only from sites with a particular set of opinions.

We can view this browsing of political webpages as a random walk on a graph, which lends itself to representation as a Markov process. Each webpage can be represented by a symbol from a finite alphabet, depending on the topic of interest. For example, we can have a 1 if the webpage

contains some number of keywords and a 0 otherwise. The transition probabilities of the process are dependent on the opinions of a particular user. Also note that we can reasonably assume that the more browser history we have, the less the additional information provided by obtaining another page of history.

Suppose we wish to quantify the polarization of opinions or information on a topic. More polarized user opinions equate to a slower mixing process. Therefore, to describe the polarization, we must obtain the transition and stationary probabilities of the process.

Given some user's browser history, we are interested in the probability the user will click on certain links. We do not know the memory of the process, i.e., how many past pages affect the click probabilities. Therefore, what we can do is to pick a context length and make transition probability estimates given contexts of that length. As the process is slow mixing, it is likely we will only be able to obtain estimates for some of the browsing contexts.

Chapter 2

Background

In this section we will first discuss the original jackknife and bootstrap procedures for i.i.d. samples. We also briefly mention cross validation. Then we will describe Markov processes. Finally, we will examine some of the prior work that has been conducted on applying resampling procedures when samples are not i.i.d.

2.1 Jackknife

2.1.1 The Bias Estimate

Consider i.i.d. samples $X_1, X_2, \ldots, X_n$ generated from an unknown source p and an estimate $\hat{\theta}(X_1^n)$ of θ, some real valued parameter of the source. The bias of an estimator,

$$\text{Bias} = E[\hat{\theta}(X_1^n) - \theta],$$

is one measure of how good the estimator is. In words, the bias tells us the difference between the expected value of the estimate and the true value of the parameter.

Without knowing the distribution from which the samples were generated, the actual bias cannot be calculated. However, the jackknife procedure [2], which was introduced by Quenouille, can be used to estimate the bias. The procedure operates by sequentially removing samples and

applying the estimator to obtain the recomputed statistics,

$$\hat{\theta}_i = \hat{\theta}(X_1, \ldots X_{i-1}, X_{i+1}, \ldots, X_n).$$

Then the bias estimate is

$$\widehat{\text{Bias}}_{\text{Jack}} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}), \tag{2.1}$$

where $\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$ is the average of the recomputed statistics. For the bias corrected estimate we have

$$\tilde{\theta} = \hat{\theta} - \widehat{\text{Bias}}_{\text{Jack}} = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)}. \tag{2.2}$$

The estimate (2.2) is not necessarily unbiased, as $\widehat{\text{Bias}}_{\text{Jack}}$ is only an estimate of the true bias of $\hat{\theta}$. In general, the correction reduces the bias from $O(\frac{1}{n})$ to $O(\frac{1}{n^2})$. If the estimator is a quadratic functional, i.e., it can be written in the form

$$\hat{\theta} = \mu^{(n)} + \frac{1}{n}\sum_{i=1}^{n}\alpha^{(n)}(x_i) + \frac{1}{n^2}\sum_{1\le i_1\le i_2\le n}\beta(x_{i_1}, x_{i_2}), \tag{2.3}$$

then $\widehat{\text{Bias}}_{\text{Jack}}$ is unbiased in estimating the true bias of the estimator [2].

Examples 1 and 2, borrowed from [2], demonstrate the jackknife procedure performed on some common estimates.

Example 1 (The Variance) Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples drawn from some distribution. The sample mean is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$ and the sample variance is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Then $\hat{\theta}_i = \frac{1}{n-1}\sum_{j\neq i}(X_j - \bar{X}_i)^2$ and $\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_i\hat{\theta}_i$, where $\bar{X}_i = \frac{1}{n-1}\sum_{j\neq i}X_j$. Applying (2.1) and (2.2) gives bias estimate

$$
\begin{aligned}
\widehat{\text{Bias}}_{\text{Jack}} &= (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}) \\
&= (n-1)\Big(\frac{1}{n}\sum_i \Big(\frac{1}{n-1}\sum_{j\neq i}(X_j - \bar{X}_i)^2\Big) - \frac{1}{n}\sum_i (X_i - \bar{X})^2\Big) \\
&= \frac{n-1}{n}\sum_i X_i^2 - \frac{2}{n}\sum_i \Big(\Big(\frac{1}{n-1}\sum_{j\neq i} X_j\Big)\sum_{j\neq i} X_j\Big) + \frac{n-1}{n}\sum_i \Big(\frac{1}{n-1}\sum_{j\neq i} X_j\Big)^2 \\
&\quad - \frac{n-1}{n}\sum_i X_i^2 + \frac{2(n-1)}{n}\Big(\frac{1}{n}\sum_i X_i\Big)\sum_i X_i - \frac{n-1}{n}\sum_i \Big(\frac{1}{n}\sum_j X_j\Big)^2 \\
&= -\frac{1}{n(n-1)}\sum_i \Big(\sum_{j\neq i} X_j\Big)^2 + \frac{n-1}{n^2}\Big(\sum_i X_i\Big)^2 \\
&= \frac{1}{n^2(n-1)}\sum_i\sum_j X_i X_j - \frac{1}{n(n-1)}\sum_i X_i^2 \\
&= -\frac{1}{n(n-1)}\sum_i (X_i - \bar{X})^2,
\end{aligned}
$$

and the bias corrected estimate is

$$
\tilde{\theta} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,
$$

which is the well known unbiased estimate of the variance. □

2.1.2 The Variance Estimate

Another measure of the quality of an estimator is its variance,

$$
\text{Var} = E[(\hat{\theta} - E\hat{\theta})^2].
$$

Tukey expanded on the jackknife procedure, using the recomputed statistics to obtain an estimate of the variance of the estimator with

$$
\widehat{\text{Var}}_{\text{Jack}} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\theta}_{(\cdot)})^2. \tag{2.4}
$$

Example 2 (The Expectation) Let $X_1, X_2, \ldots, X_n$ be as in Example 1. Here the statistic of interest is $\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\hat{\theta}_i = \frac{n\hat{\theta} - X_i}{n-1}$, $\hat{\theta}_{(\cdot)} = \hat{\theta}$, and $\hat{\theta}_i - \hat{\theta}_{(\cdot)} = \frac{\bar{X} - X_i}{n-1}$. Therefore, from

(2.4),

$$\widehat{\mathrm{Var}}_{\mathrm{Jack}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n(n-1)}.$$

□

The variance estimate can be thought of as an estimate of the variance of the statistic on a sample of size $n-1$ and then a scaling to sample size n. Let Var_{n-1} be the variance of the statistic on a sample of size $n-1$, and let

$$\widetilde{\mathrm{Var}} = \sum_{i=1}^{n}(\hat{\theta}_i - \hat{\theta}_{(\cdot)})^2$$

be the estimate of Var_{n-1}. Then

$$\widehat{\mathrm{Var}}_{\mathrm{Jack}} = \frac{n-1}{n}\widetilde{\mathrm{Var}}.$$

The main result from [3] is

$$E[\widetilde{\mathrm{Var}}] \geq \mathrm{Var_{n-1}},$$

that the variance estimate for sample size $n-1$ is biased upwards in general. More specifically, if θ is a linear functional, i.e., it can be expressed as $\hat{\theta} = \mu + \frac{1}{n}\sum_i \alpha(X_i)$, then $E[\widetilde{\mathrm{Var}}] = \mathrm{Var_{n-1}}$. This is shown via the ANOVA decomposition which splits θ as

$$\theta = \mu + \sum_i A_i(X_i) + \sum_{i \leq i'} B_{ii'}(X_i, X_{i'}) + \sum_{i \leq i' \leq i''} C_{i,i',i''}(X_i, X_{i'}, X_{i''}) + \cdots + H(X_1, X_2, \ldots X_n),$$

where $\mu = E[\theta]$ is called the grand mean, $A_i(x_i) = E[\theta|X_i = x_i] - \mu$ the ith main effect, $B_{x_i,x_{i'}} = E[\theta|X_i = x_i, X_{i'} = x_{i'}] - E[\theta|X_i = x_i] - E[\theta|X_{i'} = x_{i'}] + \mu$ the ii'th second order interaction, etc. The A_i's, $B_{ii'}$'s, etc., have mean zero and are uncorrelated. Then the variance can be written as

$$\mathrm{Var}(\theta(X_1^n)) = \frac{\sigma_\alpha^2}{n} + \binom{n-1}{1}\frac{\sigma_\beta^2}{2n^3} + \binom{n-1}{2}\frac{\sigma_\gamma^2}{3n^5} + \cdots + \frac{\sigma_\eta^2}{n^{2n}},$$

where $\sigma_\alpha^2 = \mathrm{Var}(nA(X_i))$, $\sigma_\beta^2 = \mathrm{Var}(n^2B(X_i, X_{i'}))$, $\sigma_\gamma^2 = \mathrm{Var}(n^3C(X_i, X_{i'}, X_{i''}))$, etc. A similar form can be obtained for $E[\widetilde{\mathrm{Var}}]$, and in the bias the $\frac{1}{n}$ term is cancelled out, while the result is positive.

2.2 Bootstrap

The bootstrap, introduced by Efron in [4], is another popular resampling procedure. Again consider samples $X_1, X_2, \ldots, X_n$ drawn i.i.d. from some unknown distribution p and an estimate $\hat{\theta}(X_1^n)$ of parameter θ. Now suppose we wish to estmate the variance of $\hat{\theta}(X_1^n)$. Let $\hat{p}$ be the empirical distribution given by X_1^n. Then the bootstrap estimate of the variance is obtained by simply evaluating the variance of the estimator on $\hat{p}$. Depending on the estimator, it may not be possible to explicitly calculate the variance, and so the Monte Carlo algorithm is used. To do this, begin by sampling (with replacement) from the empirical distribution to obtain i.i.d. $X_1^{*b}, X_2^{*b}, \ldots, X_n^{*b}$. Calculate the estimate on the bootstrap sample, $\hat{\theta}^{*b} = \hat{\theta}(X_1^{*b}, X_2^{*b}, \ldots, X_n^{*b})$. Repeat the process to obtain B bootstrap replications $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$. The bootstrapped estimate of the variance is then,

$$\widehat{\mathrm{Var}}_{\mathrm{Boot}} = \frac{\sum_{b=1}^{B}(\hat{\theta}^{*b} - \hat{\theta}^{*\bullet})^2}{B-1} \tag{2.5}$$

where $\hat{\theta}^{*\bullet} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$.

The following example adapted from [2], illustrates the bootstrap procedure for the variance of the average.

Example 3 (The Average) Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples from a distribution with mean μ and variance σ^2, and let $\hat{\theta}(X_1^n) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. In this case we know that $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$. Therefore the bootstrap variance estimate is $\widehat{\mathrm{Var}}_{\mathrm{Boot}} = \frac{1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$. □

The procedure can be applied to obtain estimates of other functions of θ as well. For example, to estimate $Pr(\theta \leq c)$ for some constant c, we have $Pr(\hat{\theta}^* \leq c) = \frac{1}{B}\sum_{b=1}^{B} I\{\hat{\theta}^{*b} \leq c\}$, where I is the indicator function.

For the bias, the bootstrap estimate is

$$\widehat{\mathrm{Bias}}_{\mathrm{Boot}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b} - \hat{\theta} = \hat{\theta}^{*\bullet} - \hat{\theta}. \tag{2.6}$$

In [2] it is shown that if $\hat{\theta}$ is a quadratic functional, see (2.3), then

$$\widehat{\mathrm{Bias}}_{\mathrm{Boot}} = \frac{n-1}{n}\widehat{\mathrm{Bias}}_{\mathrm{Jack}}.$$

Variations of the bootstrap method include drawing the bootstrap samples from a parametric model, e.g., a Gaussian distribution with the sample mean and variance, or using a smoothed empirical distribution obtained by convolving with a scaled normal distribution.

2.3 Cross-Validation

Cross-validation is another well known technique which falls into the category of resampling procedures. See [5]. In contrast to bootstrap and jackknife, it is generally used to estimate the error of a model built on data. Typically the data is partitioned into two sets, training and validation. The first is used to train the model and the second is used to test it, giving some error. This is then repeated multiple times by taking a different partition in some systematic manner and an average error can be obtained. For example, k-fold cross-validation involves partitioning the data into k sets, training on $k-1$ of the sets, and validating on the last set, repeating k times for each of the possibilities. In leave-one-out cross-validation, the model is training on all but one of the data points, then tested on the omitted one. The processes is then repeated for each of the data points.

2.4 Markov Processes

A Markov process $p_{\mathcal{T}}$ with finite alphabet $\mathcal{A}$ is defined by a suffix-free set of states $\mathcal{T} \subset A^* = \cup_{k\geq 0}\mathcal{A}^k$ and a set of transition probabilities $p(\mathcal{T}) = \{p(a|\mathbf{s}) > 0 : a \in \mathcal{A}, \mathbf{s} \in \mathcal{T}\}$. The set $\mathcal{T}$ can be represented as a full $\mathcal{A}$-ary tree where the leaves correspond to the states. The process has memory $D = \max\{|\mathbf{s}| : \mathbf{s} \in \mathcal{T}\}$, also equal to the depth of the tree. Here we will consider binary Markov processes, so $\mathcal{A} = \{0, 1\}$.

If a Markov process is aperiodic, irreducible[1], and has finite state space, then it has a unique stationary distribution π satisfying

$$\pi = \pi P,$$

where P is the transition matrix for the Markov chain; i.e., P_{ij} is the probability of moving from state i to state j in one step.

[1]The processes considered here are aperiodic since any state $\mathbf{s} \in \mathcal{T}$ can be reached in either $|\mathbf{s}|$ or $|\mathbf{s}|+1$ steps and irreducible since $p(a|\mathbf{s}) > 0$ for all $a \in \mathcal{A}$ and $\mathbf{s} \in \mathcal{T}$.

The following provides an example of a Markov process along with the context tree and state transition diagrams. The example also serves to intuitively illustrate some of the complications associated with estimating parameters of Markov processes.

Example 4 Let $\mathcal{A} = \{0, 1\}$ and $\mathcal{T} = \{00, 01, 10, 11\}$, with $p(1|00) = \frac{\epsilon}{m}$, $p(1|01) = 1 - \epsilon$, $p(1|10) = 1 - \epsilon$, and $p(1|11) = \epsilon$. For $m > \epsilon > 0$, the model $p_{\mathcal{T}}$ represents a stationary ergodic Markov process with stationary distribution $\pi(00) = \frac{m}{m+3}$, and $\pi(01) = \pi(10) = \pi(11) = \frac{1}{m+3}$.

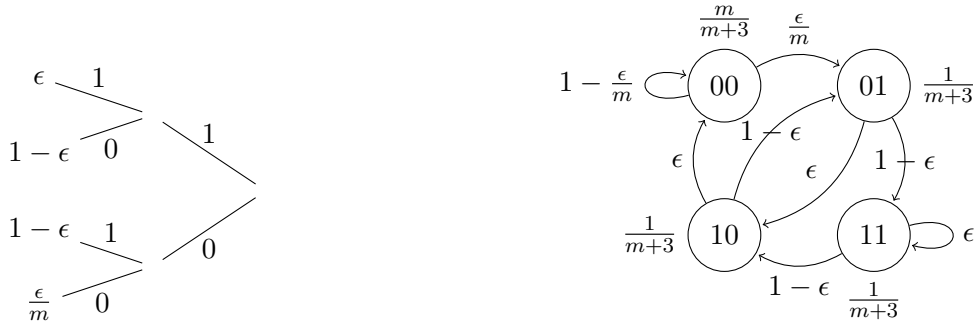
The leaves of the tree in Figure 2.1a show the conditional probability of 1 given the corresponding context. Figure 2.1b displays the state transition diagram along with the transition and stationary probabilities.

First we use this example to introduce the context tree and state transition diagrams. Suppose we had past samples $X_{-3}X_{-2}X_{-1}X_0 = 1001$. The state of the process is the suffix of the past bits that is in $\mathcal{T}$, which in this case is 01. Then $p(X_1 = 1|X_{-3}^0) = p(X_1 = 1|X_{-1}^0) = 1 - \epsilon$, since only the past at most $D = 2$ bits matter. Now suppose $X_1 = 1$. Then $X_{-3}^1 = 10011$, the current state is 11, and the next bit is a 1 with probability ϵ. Now suppose $X_2 = 0$. Then $X_{-3}^2 = 100110$, the state becomes 10, and so on.

Next we try to provide some intuition for the problems we face while making estimates with samples from a Markov process. Let $\epsilon \ll \frac{1}{n}$, where n is the sample size, and let m be large. If our process starts in state 01, 11, or 10, then with high probability our sample will look like $\ldots 011011011011 \ldots$. It is unlikely we will observe the state 00, since $p(0|10)$ is very small, even though 00 can have stationary probability which is arbitrarily close to 1. Additionally, we are not able to estimate the transition probability for 00 if it doesn't occur.

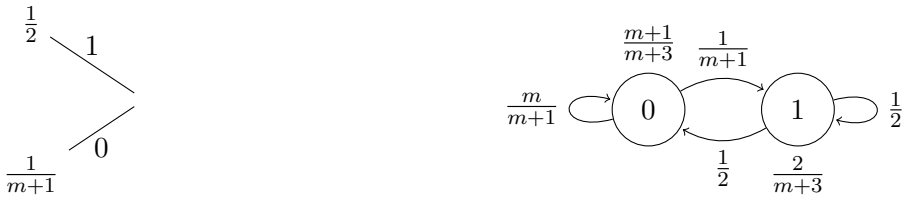
Suppose that we did not know the memory of the process or the set of states. Then we may reasonably choose to estimate $p(1|1)$ and $p(1|0)$. Based on our sample, we would guess $\frac{1}{2}$ and 1, respectively. Using the aggregate model, shown in Figure 2.2, which is discussed in the following chapter, the actual probabilities can be calculated as $p(1|1) = \frac{1}{2}$ and $p(1|0) = \frac{1}{m+1}$. We may have gotten $p(1|1)$ correct, but our estimate for $p(1|0)$ is about as far off as can be.

□



(a) (b)

Figure 2.1: (2.1a) Context tree for the Markov process from Example 4, where the leaves are the probability of 1 given the context. (2.1b) Corresponding state transition diagram with transition and stationary probabilities.



(a) (b)

Figure 2.2: (2.2a) Context tree for the aggregated process from Example 4. (2.2b) Corresponding state transition diagram with transition and stationary probabilities.

2.5 Prior Work on Resampling Methods for Markov Processes

One assumption which is essential in obtaining good results for the classical jackknife and bootstrap is that the samples are i.i.d. It is easy to see that the procedures produce poor results if the dependencies in the data are ignored. Much work has gone into adapting the jackknife and bootstrap for various situations in which the data is not i.i.d. Some of this work has focused solely on either the jackknife or the bootstrap, whle others have techniques applicable to both.

The first category of results involves grouping samples in some systematic manner. In [6] a natural extension of the jackknife is introduced, where the length-n data is partitioned into b nonoverlapping blocks of length $l = \frac{n}{b}$, and these blocks are removed one by one to obtain the recalculated statistics. It is shown that the estimator is consistent under certain conditions and also how to obtain the optimal choice for the block length. For the case that the data is from a general stationary sequence with weak dependence, [7] proposed a variation with overlapping blocks,

where the jackknife procedure omits one block at a time and the bootstrap procedure independently samples the $n - l + 1$ blocks. Here the jackknife estimates are shown to be consistent if the block length l tends to infinity at an appropriate rate. In [8], the same overlapping block procedure was independently proposed for samples which are from a stationary m-dependent[2] process. A further extension, the blocks of blocks scheme, was proposed in [9], [10], [11]. Their aim was to allow for estimators involving parameters of the entire joint distribution, e.g., the spectral density. The procedure in [12] takes a slightly different approach to selecting blocks of samples. They randomly select an initial sample, then with probability p pick the next sequential sample and with probability $1 - p$ randomly select any of the n samples. They then show the pseudo time series with geometrically distributed block lengths generated from this procedure is also stationary.

Another category of results involve estimating the transition density from the samples in order to generate more data. The Markov conditional bootstrap is introduced in [13], which estimates the transition density nonparametrically and then samples from the estimated density. It is also shown that this improves upon the block bootstrap, albeit under stronger conditions. Implicit estimation of the one step transition distribution is performed in [14], where samples from a stationary Markov process are sampled locally in order to reproduce the dependence properties of the process. The sampling procedure selects the next sample randomly from the set of values X_{s+1} such that $|X_t - X_s| \leq b$, where X_t is the current sample. [15] imposes conditional moment restrictions on the process and performs a weighted empirical estimate of the one step conditional distribution. In [16], the source is approximated by a family of parametric models, one of which is then chosen based on the data.

All of these results generally depend upon implicit or explicit mixing assumptions to obtain good estimates.

[2] A process $\{X_i\}$, $i = 1, \ldots, n$ is m-dependent if for any pair of events A and B, where A depends on $\{X_1, \ldots, X_k\}$ and B depends on $\{X_{k+m+1}, X_{k+m+2}, \ldots\}$, A and B are independent.

Chapter 3

Slow Mixing Setup

In this chapter we detail some of the potential problems encountered when making estimates with unbounded memory and no assumptions on the mixing. Next we describe our continuity condition, which allows us to be able to make reasonable estimates. Then we discuss some results from [1] involving the estimation of transition probabilities from processes that satisfy a continuity condition. We also show some simulation results of estimates made in a slow mixing process to illustrate some interesting artifacts.

3.1 Aggregated Model

Suppose we have samples $X_1, X_2, \ldots, X_n$ from a binary, ergodic, aperiodic Markov process $p_{\mathcal{T}}$. We can represent the state space of the process with the leaves of a full binary tree $\mathcal{T}$, which we assume has finite depth. Using only the samples, we wish to estimate parameters of $p_{\mathcal{T}}$.

As we have no knowledge of the actual memory or set of states, we approximate the source with an aggregated model $p_{\tilde{\mathcal{T}}}$, which has state space $\tilde{\mathcal{T}} = \{0,1\}^{k_n}$. We let k_n scale logarithmically with n for consistency, as in [17]. Now we can make estimates for parameters of the aggregate source. Let $\mathbf{s} \in \mathcal{T}$ be a state of the original process and let $\mathbf{w} \in \tilde{\mathcal{T}}$ be a state of the aggregated process. To denote that $\mathbf{w}$ is a suffix of $\mathbf{s}$, we write $\mathbf{w} \prec \mathbf{s}$.

Note that, in general, $\mathbf{w}$ is not an a state of $p_{\mathcal{T}}$, since we have no knowledge of the memory. Then for $a \in \{0, 1\}$ we have

$$p(\mathbf{w}) = \sum_{\mathbf{s}' \in \mathcal{T}, \mathbf{w} \prec \mathbf{s}'} p(\mathbf{s}')$$

and

$$p(a|\mathbf{w}) = \frac{1}{p(\mathbf{w})} \sum_{\mathbf{s} \in \mathcal{T}, \mathbf{w} \prec \mathbf{s}} p(\mathbf{s}) p(a|\mathbf{s})$$

for the aggregated stationary and conditional probabilities, respectively.

Here we give an example of calculating the probabilities for an aggregated process.

Example 5 Let $p_{\mathcal{T}}$ be a Markov process with $\mathcal{A} = \{0, 1\}$ and $\mathcal{T} = \{00, 10, 1\}$, where $p(1|00) = \frac{1}{4}$, $p(1|10) = \frac{1}{2}$, and $p(1|1) = \frac{1}{2}$. Figure 3.1 displays the context tree where the branches correspond to the context and the leaves give the probability of 1 given the context. We have that $\pi(10) = \frac{1}{5}$ and $\pi(00) = \pi(1) = \frac{2}{5}$. Let $p_{\tilde{\mathcal{T}}}$ be the aggregated process with $\tilde{\mathcal{T}} = \{0, 1\}$. Then $\tilde{p}(1|1) = \frac{1}{2}$ and $\tilde{p}(1|0) = (\frac{1}{2}\frac{1}{5} + \frac{1}{4}\frac{2}{5})/(\frac{1}{5} + \frac{2}{5}) = \frac{1}{3}$. □

(a) (b)

Figure 3.1: (3.1a) States and transition probabilities for Markov process from Example 5. (3.1b) Aggregated process with $k = 1$.

3.2 Samples From the True vs. Aggregate Source

We would like to emphasize that we do not actually see samples from the aggregated source $p_{\tilde{\mathcal{T}}}$. Our samples come from the original source $p_{\mathcal{T}}$ and we are trying to use those to estimate the parameters of the aggregated source. It is important to note this distinction, because it is possible that our original source is slow mixing, but by aggregating the states, we obtain model which is fast mixing. Then if $p_{\mathcal{T}}$ is sufficiently slow mixing relative to the sample size, samples from each process will look different. In a sample from $p_{\tilde{\mathcal{T}}}$, the empirical frequencies of states will very likely

be close to the stationary distribution, and estimates will be good in general. However, samples from $p_{\mathcal{T}}$ will reflect only a subset of the state space and estimates may not be as good. Example 6 provides an illustration.

Example 6 Let $p_{\mathcal{T}}$ be a Markov process with $\mathcal{A} = \{0, 1\}$ and $\mathcal{T} = \{0, 01, 11\}$, where $p(1|11) = 1 - \epsilon$, $p(1|01) = \epsilon$, and $p(1|0) = \frac{1}{4}$. Figure 3.2 depicts the original and aggregated processes as context trees and state transition diagrams. Calculating the stationary distribution of $p_{\mathcal{T}}$ yields $\pi(11) = \pi(01) = \frac{1}{6}$ and $\pi(0) = \frac{2}{3}$. Then the aggregated process $p_{\tilde{\mathcal{T}}}$, where $\tilde{\mathcal{T}} = \{0, 1\}$, has $\tilde{p}(1|0) = \frac{1}{4}$ and $\tilde{p}(1|1) = (\epsilon\frac{1}{6} + (1-\epsilon)\frac{1}{6})/(\frac{1}{6} + \frac{1}{6}) = \frac{1}{2}$. If $\epsilon \ll \frac{1}{n}$, then with high probability a sample from $p_{\mathcal{T}}$ will show only a subset of the states. However, the aggregate processes is fast mixing and therefore samples from each will look different.

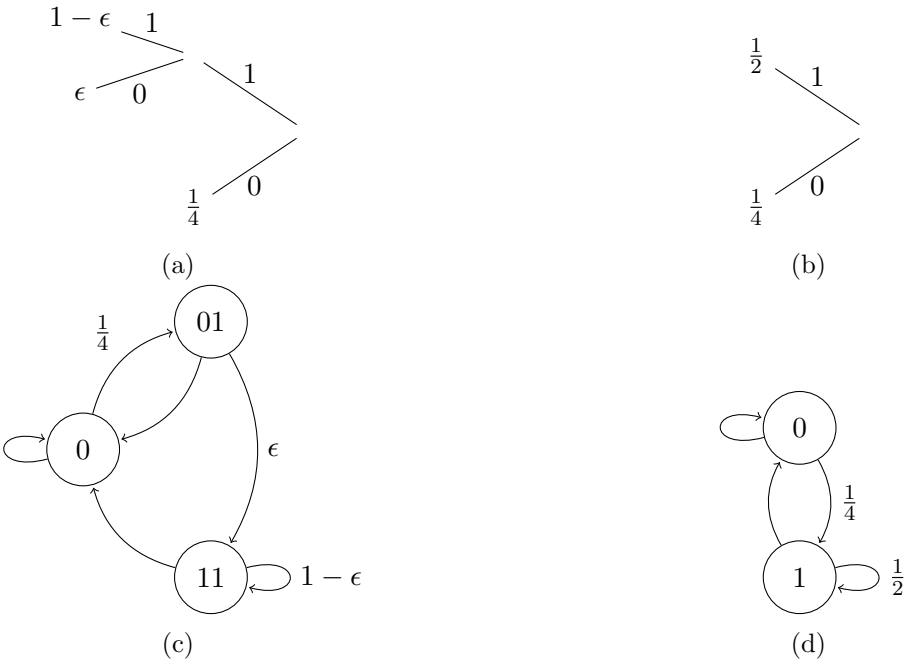


Figure 3.2: (3.2a) States and transition probabilities for Markov process from Example 6. (3.2b) Aggregated process with $k = 1$. (3.2c) State transition diagram for Markov process from Example 6. (3.2d) Aggregated process with $k = 1$.

□

To further illustrate, we simulate the processes from Example 6 and generate samples for $n = 1000$ and $\epsilon = 10^{-5}$. Let $\#\mathbf{w}$ ($\#\mathbf{w}a$) denote the number of occurrences of the string $\mathbf{w}$ ($\mathbf{w}a$). Tables

Table 3.1: Empirical frequencies and stationary probabilities of original process for Example 6.

$\mathbf{w}$	$\#\mathbf{w}/n$	$\pi(\mathbf{w})$	$\#\mathbf{w}1/\#\mathbf{w}$	$p(1\|\mathbf{w})$
0	0.805	0.667	0.243	0.25
01	0.195	0.167	0.0	0.00001
11	0.0	1.67	-	0.99999

Table 3.2: Empirical frequencies and stationary probabilities of aggregate process for Example 6.

$\mathbf{w}$	$\#\mathbf{w}/n$	$\pi(\mathbf{w})$	$\#\mathbf{w}1/\#\mathbf{w}$	$p(1\|\mathbf{w})$
0	0.672	0.667	0.239	0.25
1	0.328	0.333	0.509	0.5

3.1 and 3.2 compare the empirical frequencies with the stationary probabilities for the true and aggregate processes. Also displayed are the transition probabilities and their estimates, which will be discussed further in the following sections. In the sample from the original process, the empirical frequencies are not very close to the stationary probabilities and we only see the states 0 and 01. On the other hand, in the sample from the aggregate process, the empirical frequencies are close to their respective stationary probabilities. Additionally, note that in this sample we see occurrences of 0, 01, and 11, which is something that did not happen in the original slow mixing process.

3.3 Difficulties in Estimation

The following examples serve to illustrate some of the problems encountered when estimating processes with memory without making assumptions on the mixing.

Regardless of sample size, it is possible to have states with high stationary weight that do not appear in a sample and for the sample to be composed of strings that have very small stationary weight.

Example 7 Let $\mathcal{A} = \{0, 1\}$ and $\mathcal{T} = \{00, 01, 10, 11\}$, with $p(1|00) = \frac{\epsilon}{m}$, $p(1|01) = 1 - \epsilon$, $p(1|10) = 1 - \epsilon$, and $p(1|11) = \epsilon$. Figure 3.3 shows the state transition diagram for the process. For $m > \epsilon > 0$, the model $p_{\mathcal{T}}$ represents a stationary ergodic Markov process with stationary distribution $\pi(00) = \frac{m}{m+3}$ and $\pi(01) = \pi(10) = \pi(11) = \frac{1}{m+3}$. If m is large, then the stationary mass of 00 can be made arbitrarily close to 1 and the stationary mass of the other states arbitrarily

close to 0. Suppose we have a length-n sample and $\epsilon \ll \frac{1}{n}$. If the process starts in one of 01, 10, or 11, then with high probability we will not encounter two or more consecutive zeros. □

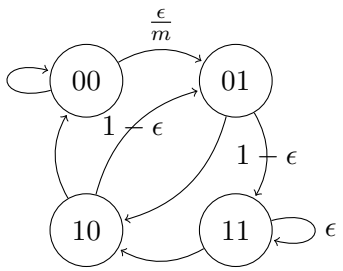


Figure 3.3: State transition diagram for Markov process from Example 7.

Note that the process in Example 7 is an example of a slow mixing process. A large number of transitions is needed before the empirical frequncies reflect the stationary probabilities because transitions between certain subsets of states occur with low probability.

This next example from [18] demonstrates that when the memory is unbounded, it may be impossible to distinguish between samples from an i.i.d. process and one with memory.

Example 8 Let $p_{\mathcal{T}}$ be a Markov process with memory $k > n$ and let $S^{(n)}$ be the set of states that can be reached in n steps when starting from the all zero state. In other words, $S^{(n)}$ is the set of all states that start with n 0's. Now let $p(1|\mathbf{s}) = \frac{1}{2}$ for all $\mathbf{s} \in S^{(n)}$ and $p(1|\mathbf{s}) = 1 - \epsilon$ for all $\mathbf{s} \notin S^{(n)}$. Then a length-$n$ sample obtained by starting from the all zero state will be indistinguishable from a length-n sample generated by an i.i.d. Bernoulli $\frac{1}{2}$ process. □

The following example illustrates some of the difficulties encountered as a result of the mixing properties of the source.

Example 9 Let $p_{\mathcal{T}}$ be as in Example 7. Now let $\mathcal{T}' = \{00, 01, 10, 11\}$, with $p'(1|00) = \epsilon$, $p'(1|01) = 1 - \epsilon$, $p'(1|10) = 1 - \epsilon$, and $p'(1|11) = \epsilon$. The corresponding state transition is shown in Figure 3.4. For $m > \epsilon > 0$, this model represents a stationary ergodic Markov process with stationary distribution $\pi'(00) = \pi'(01) = \pi'(10) = \pi'(11) = \frac{1}{4}$. Recall from Example 7 that $\pi(00) = \frac{m}{m+3}$ and $\pi(01) = \pi(10) = \pi(11) = \frac{1}{m+3}$, which varies significantly from π' when m is large. Suppose we have a length-n sample and $\epsilon \ll \frac{1}{n}$. If either process starts at 00, then with high probability, we will see n 0's. Similarly, if either process starts in 01, 10, or 11, then with high

probability we will see a length-n string like $\ldots 011011011\ldots$. In either case, the samples from both processes look identical, which means that no estimator would be able to distinguish between the two sources or obtain the correct stationary probabilities.

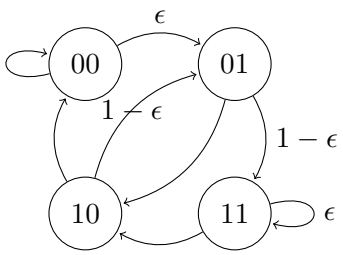


Figure 3.4: State transition diagram for Markov process from Example 9.

□

Further examples regarding the estimation of Markov processes in the slow mixing setting can be found in [1].

3.4 Conditional Probability Estimate

In what follows we consider estimating the transition probabilities $p(a|\mathbf{w})$ in a few simple situations.

3.4.1 Memory

One scenario in which we can obtain good estimates for some transition probabilities is if we know that the memory of the underlying process is less than some D. Then the subsequence following any length-D or longer context $\mathbf{w}$ is i.i.d. and $\hat{p}(a|\mathbf{w}) = \frac{\#\mathbf{w}a}{\#\mathbf{w}}$ is a reasonable estimate. Since there is no assumption on the mixing, it is possible that the state space has not been fully explored and therefore we can only provide estimates for states that have appeared sufficiently many times.

3.4.2 Mixing

On the other hand, if we had no knowledge of the memory, but we knew that the process had mixed, then we would still be able to obtain good estimates. Since the empirical counts of the strings reflect their stationary weights, then $\frac{\#\mathbf{w}a}{n}$ and $\frac{\#\mathbf{w}}{n}$ are reasonable estimates of $p(\mathbf{w}a)$ and

$p(\mathbf{w})$, respectively, and $\hat{p}(a|\mathbf{w}) = \frac{\#\mathbf{w}a}{\#\mathbf{w}}$ is a reasonable estimate of $p(a|\mathbf{w}) = \frac{p(\mathbf{w}a)}{p(\mathbf{w})}$. With the additional information of how close the stationary probabilities and empirical frequencies are, we could also bound the accuracy of the estimate.

3.4.3 Continuity Condition

Without any assumptions on the memory or the mixing of the source, the problem is ill-posed and it is not possible to say anything about our estimates. To be able to proceed further, we adopt a continuity condition as in [17] and [1]. Let $d\colon \mathbb{N}^+ \to \mathbb{R}^+$ be the continuity condition and let $\mathcal{M}_d$ be the set of all models that satisfy for all $\mathbf{u} \in \{0,1\}^*$ and $a, b, b' \in \{0,1\}$,

$$\left| \frac{p(a|b\mathbf{u})}{p(a|b'\mathbf{u})} - 1 \right| \leq d(|\mathbf{u}|). \tag{3.1}$$

Just as in [1] we require that

$$\sum_{i \geq 1} d(i) < \infty. \tag{3.2}$$

What (3.1) says is that the ratio of conditional probabilities given contexts that differ only in the least recent bit are bounded below by $1 - d(i)$ and above by $1 + d(i)$, where i is the length of the matching part of the context. In other words, $d(i)$ controls how much the last bit of a context of length $i+1$ can affect the conditional probability. If $d(i) = 0$ for all $i > D$, then we can see that the conditional probabilities for all contexts that differ only after the first D bits are equal and the process has memory D.

Requiring (3.2) means that as context length increases, the amount that the last bit can influence the conditional probability decreases. So the continuity condition can be thought of as a soft constraint on the memory of the process. From an information theoretic perspective, we are requiring that the conditional mutual information $I(X_0; X_{-i}|X_{-(i-1)}^{-1})$ decreases as i increases. Equivalently, two bits, given all the bits between, say less about each other the farther apart they are. The following example from [1] illustrates that this condition does not at all constrain the mixing of the process.

Example 10 Let $\mathcal{A} = \{0,1\}$ and $\mathcal{T} = \{0,1\}$, with $p(1|0) = \epsilon$ and $p(1|1) = 1 - \epsilon$. Here the stationary distribution is $\pi(0) = \pi(1) = \frac{1}{2}$. We can have the strong restriction that $d(i) = 0$ for all

$i > 1$, since our process has only one bit of memory. However, ϵ can be chosen arbitrarily small to make the process slow mixing. $\square$

3.5 Bound on the Conditional Probability Estimate

Given a sample $X_1, X_2, \ldots, X_n$ obtained from an unknown source $p_{\mathcal{T}}$ belonging to the class $\mathcal{M}_d$, Theorem 2 from [1] provides deviation bounds on the estimates of the conditional probabilities. Let $k = |\mathbf{w}|$, where $\mathbf{w}$ is the context of interest and let $\delta_k = \sum_{i \geq k} d(i)$. With probability at least $1 - \frac{1}{2^{2^k \log n}}$,

$$\left| \frac{\#\mathbf{w}1}{\#\mathbf{w}} - p(1|\mathbf{w}) \right| \leq \sqrt{\frac{2^k \log n + n\delta_k}{\#\mathbf{w}}}. \tag{3.3}$$

The first term of the bound increases with the length of the context and reflects the intuitive difficulty of providing good estimates with longer contexts. On the other hand, notice that the second term decreases with increasing context length. This is because the aggregated state will be closer to the true state of the process and will therefore have less bias. Also, we can see that simply having a larger sample does not ensure that our estimates will be better, since the terms in the numerator also increase with sample size. Since this result does not depend on the empirical frequencies of states being close to their stationary probabilities, nothing can necessarily be said about shorter contexts from longer ones.

3.6 Slow Mixing Simulations

We construct a slow mixing Markov process and generate samples to illustrate some of the curious artifacts we previously mentioned. The binary Markov process has memory D and generates bits using

$$p(1|X_{-\infty}^{-1}) = q(X_{-3}^{-1}) + \frac{\min\{q, 1-q\} \sum_{i=1}^{D-3} (1 - 2X_{-i-3})\alpha(i)}{2 \sum_{i=1}^{D-3} \alpha(i)} \tag{3.4}$$

where $q(000) = \frac{1}{8}$, $q(001) = 1 - \epsilon$, $q(010) = 1 - \epsilon$, $q(011) = \epsilon$, $q(100) = \frac{7}{8}$, $q(101) = \epsilon$, $q(110) = \epsilon$, $q(111) = 1 - \frac{\epsilon}{10}$, and $\alpha(i) = \frac{1}{i^2}$. Figure 3.5 depicts the state transitions for the process given by just the values of q, i.e., for $D = 3$. For ϵ sufficiently small, the process is slow mixing. We set up the process this way so that all the states with the same three most recent bits (X_{-3}^{-1}) have transition

probabilities that are more similar for longer matching contexts. This allows the process to satisfy the continuity condition. The probabilities corresponding to the three most recent bits are chosen so that the probability of certain transitions can be made arbitrarily small to make the process slow mixing.

We take $D = 15$, $\epsilon = 10^{-5}$ and generate a sample of length $n = 1000$. Our resulting continuity condition remains $d(i) = O(\frac{1}{i^2})$. For contexts of length $k \in \{1, 2, 4, 7, 10\}$, we examine the empirical and theoretical stationary and transition probabilities of states that appear sufficiently many times.

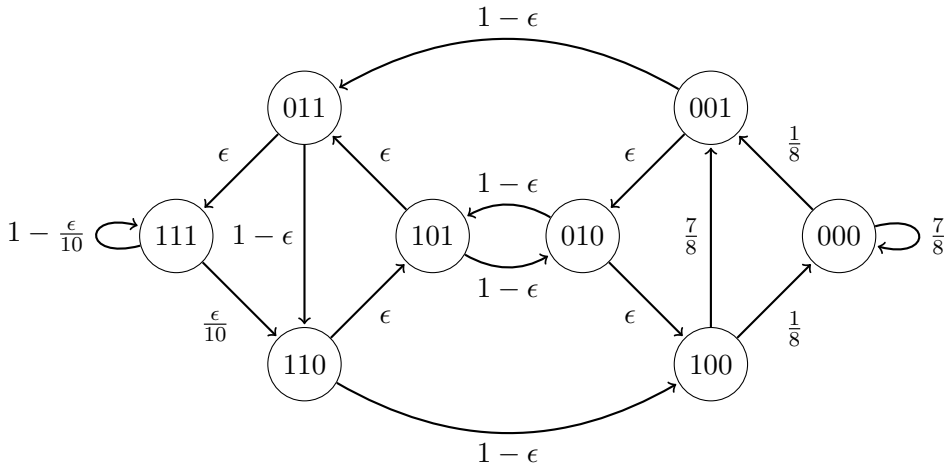


Figure 3.5: State transition diagram of base transition probabilities for slow mixing process (3.4)

Table 3.3 shows the empirical frequencies and stationary probabilities of some strings and their suffixes. Due to the slow mixing nature of the process, the empirical frequencies in general do not reflect the stationary probabilities. However, it is not uncommon to see the longer strings' counts sometimes reflect their stationary probabilities better than those of their suffixes.

The transition probability estimates and the true transition probabilities given the context $\mathbf{w}$ are displayed in Table 3.4. In some cases the estimates for longer contexts can be better than the estimates for their suffixes. While the common intuition is that it is more difficult to estimate when there is less data, this is countered by the fact that longer contexts are closer to the true state and are therefore less biased.

Recall our earlier comment about how it is important to remember that we are estimating parameters of an aggregated source with samples from the original source, since samples from the two, in general, are not similar. Here we can see that the original process given by (3.4) is slow

Table 3.3: Empirical and theoretical stationary probabilities for select states of process (3.4)

$\mathbf{w}$	$\#\mathbf{w}/n$	$\pi(\mathbf{w})$
0	0.629	0.2678
00	0.444	0.126158
0000	0.229	0.048849
0000000	0.151	0.02706
0000000000	0.095	0.014575
1100	0.184	0.068059
1001100	0.153	0.058809
0011001100	0.153	0.058806
10	0.184	0.141642
0110	0.184	0.068059
1	0.37	0.7322
01	0.185	0.141642
1001	0.154	0.058811
0011001	0.154	0.058809
0110011001	0.126	0.050727
11	0.185	0.590557
0011	0.185	0.068059

Table 3.4: Empirical and theoretical transition probabilities for select states of process (3.4)

$\mathbf{w}$	$\#\mathbf{w}1/\#\mathbf{w}$	$p(1\|\mathbf{w})$	$\#\mathbf{w}0/n$	$\#\mathbf{w}1/n$
0	0.294118	0.528911	0.444	0.185
00	0.416667	0.539479	0.259	0.185
0000	0.126638	0.172624	0.2	0.029
0000000	0.125828	0.185566	0.132	0.019
0000000000	0.147368	0.187268	0.081	0.014
1100	0.836957	0.864108	0.03	0.154
1001100	0.823529	0.862599	0.027	0.126
0011001100	0.823529	0.862599	0.027	0.126
10	0.0	0.519498	0.184	0.0
0110	0.0	0.000013	0.184	0.0
1	0.5	0.806552	0.185	0.185
01	1.0	0.480503	0.0	0.185
1001	1.0	0.999987	0.0	0.154
0011001	1.0	0.999987	0.0	0.154
0110011001	1.0	0.999987	0.0	0.126
11	0.0	0.884754	0.185	0.0
0011	0.0	0.000012	0.185	0.0

Table 3.5: Empirical and theoretical stationary and transition probabilities for select states of aggregate process

$\mathbf{w}$	$\#\mathbf{w}/n$	$\pi(\mathbf{w})$	$\#\mathbf{w}1/\#\mathbf{w}$	$p(1\|\mathbf{w})$
0	0.271	0.2678	0.542435	0.528911
00	0.125	0.126158	0.528	0.539479
10	0.146	0.141642	0.554794	0.519498
1	0.729	0.7322	0.798353	0.806552
01	0.147	0.141642	0.476190	0.480503
11	0.582	0.590557	0.879725	0.884754

mixing, but the aggregated model with memory $k = 2$ is not. To show how samples from the two processes can differ, we generate a length $n = 1000$ sample from the memory-2 aggregate source and compare the estimates. Table 3.5 displays the estimates of the stationary and transition probabilities for the sample from the aggregate process, along with the true values. Observe that the estimates of both the stationary and transition probabilities are close to their true values, whereas in the original process, this is not the case.

Chapter 4

The Jackknife Variance Estimate

We apply a modification of the jackknife procedure to obtain an estimate of the variance of the conditional probability estimates and give bounds on the bias of the jackknife variance estimate, which we first presented in [18]. The theorem in this chapter gives a slightly improved bound.

4.1 Jackknife Procedure for the Variance of $\frac{\#\mathbf{w}1}{\#\mathbf{w}}$

Here we describe a natural adaptation to the classical jackknife procedure for estimating the variance when our statistic of interest is the conditional probability of the aggregated process given some context.

Let $Y_1, Y_2, \ldots, Y_m$ be the subsequence following the context $\mathbf{w}$, where m is the number of times $\mathbf{w}$ appears in the sample. From [1], we know that it is reasonable to estimate $p(1|\mathbf{w})$ with

$$\hat{Y} = \frac{Y_1 + Y_2 + \cdots + Y_m}{m}. \tag{4.1}$$

Note that the mean $\hat{Y}$ is the same as $\frac{\#\mathbf{w}1}{\#\mathbf{w}}$. Then in the same manner as the i.i.d. jackknife, we sequentially remove samples to obtain the recalculated means,

$$\hat{Y}_i = \frac{Y_1 + \cdots + Y_{i-1} + Y_{i+1} + \cdots + Y_m}{m-1}. \tag{4.2}$$

Let $\hat{Y}_{(\cdot)} = \frac{1}{m}\sum_{i=1}^{m}\hat{Y}_i$ be the average of the recalculated means. Then the jackknife estimate of $\mathrm{Var}(\hat{Y})$ is

$$\mathrm{Var}_{\mathrm{Jack}} = \frac{m-1}{m}\sum_{i=1}^{m}(\hat{Y}_i - \hat{Y}_{(\cdot)})^2. \tag{4.3}$$

Remark Note that for everything that follows, $\mathrm{Var}_{\mathrm{Jack}}$ denotes the jackknife estimate of the variance of $\frac{\#\mathbf{w}1}{\#\mathbf{w}}$, the empirical estimate of the probability of 1 given the length-k context of consideration $\mathbf{w}$. As we do not deviate from the above setting, we drop further references to $\mathbf{w}$. □

Proposition 1

$$\mathrm{Var}_{\mathrm{Jack}} = \frac{1}{m(m-1)}\sum_{i=1}^{m}(Y_i - \hat{Y})^2. \tag{4.4}$$

Proof Rewriting in terms of Y_j's and simplifying, we can see that

$$\hat{Y}_i - \hat{Y}_{(\cdot)} = \frac{\sum_{j\neq i}Y_j}{m-1} - \frac{\sum_{i'=1}^{m}\sum_{j\neq i'}Y_j}{m(m-1)} = \frac{-mY_i + \sum_{j=1}^{m}Y_j}{m(m-1)} = \frac{\hat{Y} - Y_i}{m-1},$$

and inserting into (4.3) gives the result. □

4.2 Bias of the Jackknife Estimate

Here we show how the equation for the jackknife estimate of the variance in Proposition 1 can be obtained by applying the bias correction (2.2) to the naive empirical variance estimate. We have our statistic

$$\hat{\theta}(Y_1^m) = \frac{\sum_{i=1}^{m}(Y_i - \hat{Y})^2}{m},$$

so $\hat{\theta}_i = \frac{1}{m-1}\sum_{j\neq i}(Y_j - \hat{Y}_i)^2$ and $\hat{\theta}_{(\cdot)} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$, where $\hat{Y}_i$ is as in (4.2). Then calculating the bias corrected estimate gives the unbiased estimate for the variance of Y_i,

$$m\hat{\theta}(Y_1^m) - (m-1)\hat{\theta}_{(\cdot)} = \frac{\sum_{i=1}^{m}(Y_i - \hat{Y})^2}{m-1}.$$

To obtain the variance estimate of $\hat{Y}$ for i.i.d. Y_i's see that

$$\begin{aligned}
\mathrm{Var}(\hat{Y}) &= \mathrm{Var}\Big(\frac{1}{m}\sum_{i=1}^{m} Y_i\Big) \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}(Y_i) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j\neq i}\mathrm{Cov}(Y_i, Y_j) \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}(Y_i) \\
&= \frac{1}{m}\mathrm{Var}(Y_i)
\end{aligned}$$

where the last two lines follow from the independence and identical distribution of the Y_i's, respectively. Therefore, for the i.i.d. case, we have the unbiased jackknife estimate of the variance

$$\mathrm{Var}_{\mathrm{Jack}} = \frac{\sum_{i=1}^{m}(Y_i - \hat{Y})^2}{m(m-1)}.$$

In our Markov setting, we have that the Y_i's may be correlated, and the true state generating each may be different, so our estimate is not unbiased. We analyze the bias in Theorem 4 where we decompose the Y_i's into one set of random variables which form a martingale difference sequence and another which is well behaved.

4.3 Bound on the Jackknife Variance Estimate

Let $\mathcal{T}$ be the set of states of the unknown process p, and let $\mathcal{T}_{\mathbf{w}} = \{\mathbf{s} \in \mathcal{T} : \mathbf{w} \prec \mathbf{s}\}$ be the set of true states corresponding to aggregated state $\mathbf{w}$. We are examining the conditional probability for the context $\mathbf{w}$, while each bit in the subsequence is generated by some state $\mathbf{s}$ that has $\mathbf{w}$ as a suffix. Therefore, we want to bound the difference between the two probabilities, and the continuity condition allows us to do so.

Proposition 2 For $\delta_k \leq 1$ and $a \in \{0,1\}$, $\mathbf{w} \in \{0,1\}^k$, and $\mathbf{s} \in \mathcal{T}_{\mathbf{w}}$,

$$|p(a|\mathbf{w}) - p(a|\mathbf{s})| \leq \delta_k.$$

Proof From [1] we have for $\delta_k \leq 1$,

$$(1-\delta_k)\max_{\mathbf{s}\in\mathcal{T}_\mathbf{w}} p(a|\mathbf{s}) \leq p(a|\mathbf{w}) \leq \frac{1}{1-\delta_k}\min_{\mathbf{s}\in\mathcal{T}_\mathbf{w}} p(a|\mathbf{s})$$

Taking the left inequality we get $(1-\delta_k)p(a|\mathbf{s}) \leq p(a|\mathbf{w})$. Rearranging gives $p(a|\mathbf{s}) - p(a|\mathbf{w}) \leq p(a|\mathbf{s})\delta_k \leq \delta_k$, where the last inequality follows because, as a probability, $p(a|\mathbf{s}) \leq 1$. From the right side inequality we have, $p(a|\mathbf{w}) \leq \frac{p(a|\mathbf{s})}{1-\delta_k}$, and similar to the left side we can see $(1-\delta_k)p(a|\mathbf{w}) \leq p(a|\mathbf{s})$ and $p(a|\mathbf{w}) - p(a|\mathbf{s}) \leq p(a|\mathbf{w})\delta_k \leq \delta_k$. □

Definition 1 Let $1_i(s)$ be the indicator function that the state corresponding to the ith appearance of $\mathbf{w}$ is $\mathbf{s}$. Let $W_i = \sum_{\mathbf{s}\in\mathcal{T}_\mathbf{w}} 1_i(\mathbf{s})p(1|\mathbf{s})$ and let $Z_i = Y_i - W_i$. □

In other words, W_i is the probability of 1 given the true context $\mathbf{s}$ that generated Y_i. The Z_i's form a martingale difference sequence, which has the following properties.

Lemma 1 Let $\mathcal{F}_i$ denote the sigma algebra of $Z_1, \ldots, Z_i$. For $i < j$,

$$E[Z_j|\mathcal{F}_i] = 0$$

and

$$E[Z_i Z_j] = 0.$$

Proof For the first part we show that $E[Z_i \mid Z_1^{i-1}] = 0$ for all i. Let S_i be the true state of the process that corresponds to the ith appearance of $\mathbf{w}$. Then

$$\begin{aligned}
E[Z_i \mid Z_1^{i-1}] &= E_{S_i}[E[Z_i \mid Z_1^{i-1}, S_i]] \\
&= E_{S_i}[E[Y_i - W_i \mid Z_1^{i-1}, S_i]] \\
&= E_{S_i}[E[p(Y_i \mid S_i)] - p(1 \mid S_i)] \\
&= E_{S_i}[p(1 \mid S_i) - p(1 \mid S_i)] \\
&= 0.
\end{aligned}$$

The first two equalities follow by conditioning on the true state S_i and by the definition of Z_i. The third follows by the Markov property and definition of W_i and the fourth by evaluating the inner expectation.

The second part follows from conditioning on $\mathcal{F}_i$ and applying the first part,

$$E[Z_i Z_j] = E[E[Z_i Z_j \mid \mathcal{F}_i]] = E[Z_i E[Z_j \mid \mathcal{F}_i]] = E[Z_i \cdot 0] = 0.$$

□

Proposition 3 For $\delta_k \leq 1$ and all $m > 0$,

$$\left| \left(\frac{1}{m} \sum_{i=1}^{m} W_i \right) - p(1 \mid \mathbf{w}) \right| \leq \delta_k.$$

Proof We have that $\sum_{i=1}^{m} \sum_{\mathbf{s} \in \mathcal{T}_\mathbf{w}} 1_i(\mathbf{s}) = m$, since the inner summation evaluates to 1. Then $\frac{1}{m} \sum_{i=1}^{m} W_i = \frac{1}{m} \sum_{i=1}^{m} \sum_{\mathbf{s} \in \mathcal{T}_\mathbf{w}} 1_i(\mathbf{s}) p(1|\mathbf{s})$ is a convex combination of the probabilities $p(1|\mathbf{s})$, where $\mathbf{s} \in \mathcal{T}_\mathbf{w}$, so applying Proposition 2 gives us the result. □

Here we give two lemmas that will be useful in bounding the variances and covariances.

Lemma 2 For any random variable X with $|X| \leq \epsilon$,

$$\mathrm{Var}[X] \leq \epsilon^2.$$

Proof Since $|X|^2 \leq \epsilon^2$ and $\mathrm{Var}[X] = E[X^2] - E[X]^2$, we know that $0 \leq E[X^2] \leq \epsilon^2$ and $0 \leq E[X]^2 \leq \epsilon^2$. Therefore $E[X^2] - E[X]^2 \leq \epsilon^2$. □

Lemma 3 For any two random variables U and V,

$$|\mathrm{Cov}(U, V)| \leq \sqrt{\mathrm{Var}[U]\mathrm{Var}[V]}.$$

Proof By the Cauchy-Schwarz inequality,

$$\begin{aligned}|\mathrm{Cov}(U,V)| &= |E[(U-E[U])(V-E[V])]| \\ &\leq \sqrt{E[(U-E[U])^2]E[(V-E[V])^2]} \\ &= \sqrt{\mathrm{Var}[U]\mathrm{Var}[V]}.\end{aligned}$$

□

Now we give the theorem bounding the bias of the jackknife variance of $\hat{Y}$.

Theorem 4 For $\delta_k \leq 1$,

$$\left| E[\mathrm{Var}_{\mathrm{Jack}}|m] - \mathrm{Var}[\hat{Y}|m] \right| \leq \delta_k^2 + \frac{2\delta_k}{\sqrt{m}} + \frac{2\delta_k}{m-1}$$

Proof The bound comes from decomposing the Y_i's into W_i's and Z_i's, explicitly calculating the bias, and then bounding the resulting covariance terms.

Let $\hat{W} = \frac{1}{m}\sum_{i=1}^m W_i$ and $\hat{Z} = \frac{1}{m}\sum_{i=1}^m Z_i$. Define

$$\tilde{Z} = \frac{1}{m^2}\sum_{i=1}^m E[Z_i^2|m]. \tag{4.5}$$

First we have,

$$\begin{aligned}\mathrm{Var}[\hat{Y}|m] &= \mathrm{Var}[\hat{W}+\hat{Z}|m] \\ &= \mathrm{Var}[\hat{W}|m] + \mathrm{Var}[\hat{Z}|m] + 2\mathrm{Cov}(\hat{W},\hat{Z}|m).\end{aligned} \tag{4.6}$$

Now see that

$$\begin{aligned}\mathrm{Var}[\hat{Z}|m] &= \frac{1}{m^2}\Big(\sum_{i=1}^m E[Z_i^2|m] + \sum_{i=1}^m\sum_{j\neq i} E[Z_iZ_j|m] - \sum_{i=1}^m (E[Z_i|m])^2 - \sum_{i=1}^m\sum_{j\neq i} E[Z_i]E[Z_j]\Big) \\ &= \frac{1}{m^2}\sum_{i=1}^m E[Z_i^2|m] \\ &= \tilde{Z}\end{aligned} \tag{4.7}$$

where the second line follows because by Lemma 1 $E[Z_i Z_j|m] = 0$ and $E[Z_i|m] = E[E[Z_i|\mathcal{F}_j, m]] = 0$. Then for the variance of $\hat{Y}$, we have

$$\text{Var}[\hat{Y}|m] = \text{Var}[\hat{W}|m] + \tilde{Z} + 2\text{Cov}(\hat{W}, \hat{Z}|m). \tag{4.8}$$

Next, from Proposition 1 and since $Y_i - \hat{Y} = (W_i - \hat{W}) + (Z_i - \hat{Z})$, we have

$$\begin{aligned} E[\text{Var}_{\text{Jack}}|m] &= \frac{E\Big[\sum_{i=1}^{m}(Y_i - \hat{Y})^2\Big|m\Big]}{m(m-1)} \\ &= \frac{\sum_{i=1}^{m}\Big(E\Big[(W_i - \hat{W})^2\Big|m\Big] + 2E\Big[(W_i - \hat{W})(Z_i - \hat{Z})\Big|m\Big] + E\Big[(Z_i - \hat{Z})^2\Big|m\Big]\Big)}{m(m-1)} \end{aligned} \tag{4.9}$$

Taking the term with only Z_i's we can see that

$$\begin{aligned} \frac{E\left[\sum_{i=1}^{m}(Z_i - \hat{Z})^2 \mid m\right]}{m(m-1)} &= \frac{\sum_{i=1}^{m}\Big(E[Z_i^2 \mid m] - 2E[Z_i\hat{Z} \mid m]\Big)}{m(m-1)} + \frac{E[\hat{Z}^2 \mid m]}{m-1} \\ &= \frac{\sum_{i=1}^{m}\Big(E[Z_i^2|m] - \frac{2}{m}\Big(E[Z_i^2|m] + \sum_{j\neq i}E[Z_i Z_j|m]\Big)\Big)}{m(m-1)} \\ &\quad + \frac{\sum_{i=1}^{m}E[Z_i^2|m] + \sum_{i=1}^{m}\sum_{j\neq i}E[Z_i Z_j|m]}{m^2(m-1)} \\ &= \frac{\sum_{i=1}^{m}\Big((1-\frac{2}{m})E[Z_i^2|m]\Big)}{m(m-1)} + \frac{\sum_{i=1}^{m}E[Z_i^2|m]}{m^2(m-1)} \\ &= \frac{\frac{m-1}{m}\sum_{i=1}^{m}E[Z_i^2 \mid m]}{m(m-1)} \\ &= \frac{1}{m^2}\sum_{i=1}^{m}E[Z_i^2 \mid m] \qquad (4.10) \\ &= \tilde{Z} \qquad (4.11) \end{aligned}$$

where the third equality follows as a result of lemma 1. Therefore, for the expected variance estimate, we have

$$E[\text{Var}_{\text{Jack}}|m] = \frac{1}{m(m-1)}\sum_{i=1}^{m}E\Big[(W_i - \hat{W})^2\Big|m\Big] + \frac{2}{m(m-1)}\sum_{i=1}^{m}E\Big[(W_i - \hat{W})(Z_i - \hat{Z})\Big|m\Big] + \tilde{Z}. \tag{4.12}$$

Taking the difference between (4.12) and (4.6), the $\tilde{Z}$'s cancel and we are left with

$$\begin{aligned} &E[\mathrm{Var}_{\mathrm{Jack}}|m] - \mathrm{Var}[\hat{Y}|m] \\ &\quad = \frac{1}{m(m-1)} \sum_{i=1}^{m} E\Big[(W_i - \hat{W})^2 \Big| m\Big] + \frac{2}{m(m-1)} \sum_{i=1}^{m} E\Big[(W_i - \hat{W})(Z_i - \hat{Z}) \Big| m\Big] \\ &\qquad - \mathrm{Var}[\hat{W}|m] - 2\mathrm{Cov}(\hat{W}, \hat{Z}|m). \end{aligned} \tag{4.13}$$

Next we can combine the terms of (4.13) that have only W_i's. First, expanding the variance of $\hat{W}$ gives

$$\begin{aligned} \mathrm{Var}[\hat{W}|m] &= \frac{1}{m^2} \mathrm{Var}\Big[\sum_{i=1}^{m} W_i \Big| m\Big] \\ &= \frac{1}{m^2} \sum_{i=1}^{m} \mathrm{Var}[W_i|m] + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j \neq i} \mathrm{Cov}(W_i, W_j|m) \\ &= \frac{1}{m^2} \sum_{i=1}^{m} E[W_i^2|m] - \frac{1}{m^2} \sum_{i=1}^{m} \big(E[W_i|m]\big)^2 + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j \neq i} \mathrm{Cov}(W_i, W_j|m). \end{aligned} \tag{4.14}$$

Similarly for the sample variance term,

$$\begin{aligned} \frac{\sum_{i=1}^{m} E\Big[(W_i - \hat{W})^2 \Big| m\Big]}{m(m-1)} &= \frac{1}{m(m-1)} \Bigg(\sum_{i=1}^{m} E\Bigg[W_i^2 - \frac{2}{m} \sum_{j=1}^{m} W_i W_j + \frac{1}{m^2} \sum_{j=1}^{m} \sum_{k=1}^{m} W_j W_k \Bigg| m \Bigg] \Bigg) \\ &= \frac{1}{m(m-1)} \sum_{i=1}^{m} E[W_i^2|m] - \frac{1}{m^2(m-1)} \sum_{i=1}^{m} \sum_{j=1}^{m} E[W_i W_j|m] \\ &= \frac{1}{m^2} \sum_{i=1}^{m} E[W_i^2|m] - \frac{1}{m^2(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} E[W_i W_j|m]. \end{aligned} \tag{4.15}$$

Then combining the two, $E[W_i^2|m]$ cancels and we get

$$\begin{aligned}
&\mathrm{Var}[\hat{W}|m] - \frac{1}{m(m-1)}\sum_{i=1}^{m} E[(W_i - \hat{W})^2|m] \\
&\quad = -\frac{1}{m^2}\sum_{i=1}^{m}\big(E[W_i|m]\big)^2 + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j\neq i}\mathrm{Cov}(W_i, W_j|m) + \frac{1}{m^2(m-1)}\sum_{i=1}^{m}\sum_{j\neq i} E[W_i W_j|m] \\
&\quad = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j\neq i}\mathrm{Cov}(W_i, W_j|m) + \frac{1}{m^2(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}\Big(E[W_i W_j|m] - \big(E[W_i|m]\big)^2\Big) \\
&\quad = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j\neq i}\mathrm{Cov}(W_i, W_j|m) + \frac{1}{m^2(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}\mathrm{Cov}(W_i, W_j|m) \\
&\quad = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}\mathrm{Cov}(W_i, W_j|m)
\end{aligned}$$

and returning to (4.13) we have,

$$\begin{aligned}
&E[\mathrm{Var}_{\mathrm{Jack}}|m] - \mathrm{Var}[\hat{Y}|m] \\
&\quad = -\frac{\sum_{i=1}^{m}\sum_{j\neq i}\mathrm{Cov}(W_i, W_j|m)}{m(m-1)} + \frac{2\sum_{i=1}^{m} E\Big[(W_i - \hat{W})(Z_i - \hat{Z})\Big|m\Big]}{m(m-1)} - 2\mathrm{Cov}(\hat{W}, \hat{Z}|m) \qquad (4.16)
\end{aligned}$$

Now we can bound each of the terms. For the second one we can apply the Cauchy Schwarz inequality to show

$$\begin{aligned}
\left|E\left[\frac{\sum_{i=1}^{m}(W_i - \hat{W})(Z_i - \hat{Z})}{m}\Big|m\right]\right| &\leq \left|E\left[\sqrt{\frac{\sum_{i=1}^{m}(W_i - \hat{W})^2}{m}\frac{\sum_{i=1}^{m}(Z_i - \hat{Z})^2}{m}}\Big|m\right]\right| \\
&\leq \sqrt{E\Big[\frac{\sum_{i=1}^{m}(W_i - \hat{W})^2}{m}\Big|m\Big]E\Big[\frac{\sum_{i=1}^{m}(Z_i - \hat{Z})^2}{m}\Big|m\Big]} \\
&\leq \sqrt{\delta_k^2\frac{m-1}{m}} \\
&\leq \delta_k \qquad (4.17)
\end{aligned}$$

where the first inequality uses the regular form $|\sum a_i b_i| \leq \sqrt{\sum a_i^2 \sum b_i^2}$ and the second uses the expectation form $|E[AB]| \leq \sqrt{E[A^2]E[B^2]}$.

For the third inequality, see that

$$\begin{aligned} E\Big[(W_i-\hat{W})^2\Big|m\Big] &= E\Big[\Big(\big(W_i-p(1|\mathbf{w})\big)-\Big(\frac{1}{m}\sum_{j=1}^{m}\big(W_j-p(1|\mathbf{w})\big)\Big)\Big)^2\Big|m\Big] \\ &= E\Big[\big(W_i-p(1|\mathbf{w})\big)^2\Big|m\Big]-\frac{1}{m}\sum_{j=1}^{m}E\Big[\big(W_i-p(1|\mathbf{w})\big)\big(W_j-p(1|\mathbf{w})\big)\Big|m\Big] \\ &\leq \delta_k^2, \end{aligned}$$

that is, adding a constant to W_i doesn't change the sample variance. Then since $|W_i-p(1|\mathbf{w})|\leq\delta_k$ the same reasoning as in the proof of Lemma 2 can be applied. Also from (4.10), $E\big[\frac{\sum_{i=1}^{m}(Z_i-\hat{Z})^2}{m}\big|m\big]=\frac{m-1}{m^2}\sum_{i=1}^{m}E[Z_i^2|m]\leq\frac{m-1}{m}$, since $Z_i^2\leq 1$.

Now we bound the other terms. By Lemma 2 and Proposition 2,

$$\mathrm{Var}[W_i|m]=\mathrm{Var}[W_i-p(1|\mathbf{w})|m]\leq\delta_k^2. \tag{4.18}$$

Similarly, by Lemma 2 and Proposition 3,

$$\mathrm{Var}[\hat{W}|m]=\mathrm{Var}[\hat{W}-p(1|\mathbf{w})|m]\leq\delta_k^2. \tag{4.19}$$

From (4.7) and since $|Z_i|\leq 1$,

$$\mathrm{Var}[\hat{Z}|m]=\frac{1}{m^2}\sum_{i=1}^{m}E[Z_i^2|m]\leq\frac{1}{m}. \tag{4.20}$$

Then by Lemma 3 with (4.18) we have

$$|\mathrm{Cov}(W_i,W_j|m)|\leq\sqrt{\mathrm{Var}[W_i|m]\mathrm{Var}[W_j|m]}\leq\delta_k^2 \tag{4.21}$$

and with (4.19) and (4.20),

$$\Big|\mathrm{Cov}(\hat{W},\hat{Z}|m)\Big|\leq\sqrt{\mathrm{Var}[\hat{W}|m]\mathrm{Var}[\hat{Z}|m]}\leq\sqrt{\delta_k^2\frac{1}{m}}=\frac{\delta_k}{\sqrt{m}}. \tag{4.22}$$

Then applying (4.17), (4.21), and (4.22) to (4.16), we have

$$\begin{aligned}\left|E[\mathrm{Var}_{\mathrm{Jack}}|m] - \mathrm{Var}[\hat{Y}|m]\right| &\leq \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}\delta_k^2 + \frac{2}{m-1}\delta_k + 2\frac{\delta_k}{\sqrt{m}} \\ &= \delta_k^2 + \frac{2\delta_k}{m-1} + \frac{2\delta_k}{\sqrt{m}}, \end{aligned} \tag{4.23}$$

finishing the proof of the theorem.

□

Recall that $\delta_k \to 0$ as $k \to \infty$. We can see that the bias of the variance estimate decreases as the context length increases. If the context length is longer than the memory of the source, then we have $\delta_k = 0$ and the subsequence Y_1^m is i.i.d., resulting in an unbiased estimate. In contrast to the bound on the transition probability estimate (3.3), there is no term that increases with context length. However, for longer contexts, we will have smaller m, implying a larger bias. Notice that even if we let $m \to \infty$, we still have δ_k^2. This term comes from the covariance between W_i's and remains because we do not know the true state $\mathbf{s}$ and are instead using an aggregated state $\mathbf{w}$.

Chapter 5

The Bootstrap Variance Estimate

Here we apply the bootstrap procedure to estimate the variance of the conditional probability estimates. We see how the bootstrap relates to the jackknife for these estimates and obtain a bound on the bias of the bootstrap estimate.

5.1 Bootstrap Variance of $\frac{\#\mathbf{w}1}{\#\mathbf{w}}$

Recall that $Y_1, Y_2, \ldots Y_m$ is the subsequence following the context $\mathbf{w}$, and the empirical estimate of the conditional probability $p(1|\mathbf{w})$ is $\hat{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$. We want to use bootstrap to obtain an estimate of the variance of $\hat{Y}$.

We could apply the Monte Carlo algorithm to produce an estimate of the variance. Generating i.i.d. samples Y_i^{*b}, $b = 1 \ldots B$, $i = 1 \ldots m$, where Y_i^{*b} is Bernoulli with $Pr(Y_i^{*b} = 1) = \hat{Y}$, gives bootstrap replications $\hat{Y}^{*b} = \frac{1}{m}\sum_{i=1}^{m} Y_i^{*b}$ and variance estimate

$$\frac{1}{B-1}\sum_{b=1}^{B}(\hat{Y}^{*b} - \hat{Y}^{*\bullet})^2$$

where $\hat{Y}^{*\bullet} = \frac{1}{B}\sum_{b=1}^{B}\hat{Y}^{*b}$.

However, this is unnecessary as we are able to perform the calculations for the variance. Let $\mathrm{Var}_{\mathrm{Boot}}$ denote the bootstrap estimate of the variance of $\hat{Y}$. Then we have

$$\mathrm{Var}_{\mathrm{Boot}} = \frac{\hat{Y}(1-\hat{Y})}{m} \tag{5.1}$$

Note that just like in the jackknife case, the variance estimate here is based on the samples being i.i.d. As they are not, we analyze the correlations betwen Y_i's to provide a bound on the bias.

5.2 Bootstrap vs. Jackknife Estimate

We can see that the bootstrap estimate proposed above is actually a scaling of the estimate resulting from the earlier jackknife procedure,

$$\begin{aligned}
\mathrm{Var}_{\mathrm{Boot}} &= \frac{1}{m}(\hat{Y} - \hat{Y}^2) \\
&= \frac{1}{m}\Big(\frac{1}{m}\sum_{i=1}^{m} Y_i - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} Y_i Y_j\Big) \\
&= \frac{1}{m}\Big(\frac{1}{m}\sum_{i=1}^{m} Y_i - \frac{2}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} Y_i Y_j + \frac{1}{m^3}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} Y_j Y_k\Big) \\
&= \frac{1}{m^2}\sum_{i=1}^{m}(Y_i - \hat{Y})^2 \\
&= \Big(\frac{m-1}{m}\Big)^2 \sum_{i=1}^{m}(\hat{Y}_i - \hat{Y}_{(\cdot)})^2 \\
&= \frac{m-1}{m}\mathrm{Var}_{\mathrm{Jack}}.
\end{aligned}$$

This also means that in the i.i.d. case, the bootstrap variance is biased downwards,

$$E[\mathrm{Var}_{\mathrm{Boot}}] = \frac{m-1}{m}\mathrm{Var}(\hat{Y}).$$

5.3 Bound on the Bootstrap Variance Estimate

Similar to the section on jackknife, we bound the bias of the bootstrap variance for the conditional probability estimates.

Theorem 5 For $\delta_k \leq 1$,

$$\Big|E[\mathrm{Var}_{\mathrm{Boot}}|m] - \mathrm{Var}[\hat{Y}|m]\Big| \leq \delta_k^2 + \frac{2\delta_k}{\sqrt{m}} + \frac{2\delta_k}{m} + \frac{1}{m^2}.$$

Proof We take the same approach as in the proof of the jackknife bias for obtaining the bound.

From (4.6) we have

$$\mathrm{Var}[\hat{Y}|m] = \mathrm{Var}[\hat{W}|m] + \mathrm{Var}[\hat{Z}|m] + 2\mathrm{Cov}(\hat{W}, \hat{Z}|m). \tag{5.2}$$

and scaling (4.12) appropriately gives

$$\begin{aligned} E[\mathrm{Var}_{\mathrm{Boot}}|m] &= \frac{m-1}{m} E[\mathrm{Var}_{\mathrm{Jack}}|m] \\ &= \frac{1}{m^2}\sum_{i=1}^{m} E\Big[(W_i - \hat{W})^2 \Big| m\Big] + \frac{2}{m^2}\sum_{i=1}^{m} E\Big[(W_i - \hat{W})(Z_i - \hat{Z})\Big| m\Big] + \frac{m-1}{m}\tilde{Z}. \end{aligned} \tag{5.3}$$

Taking the difference of (5.3) and (5.2) gives the bias,

$$\begin{aligned} E[\mathrm{Var}_{\mathrm{Boot}}|m] - \mathrm{Var}[\hat{Y}|m] = \frac{1}{m^2}\sum_{i=1}^{m} E\Big[(W_i - \hat{W})^2 \Big| m\Big] + \frac{2}{m^2}\sum_{i=1}^{m} E\Big[(W_i - \hat{W})(Z_i - \hat{Z})\Big| m\Big] \\ - \frac{1}{m}\tilde{Z} - \mathrm{Var}[\hat{W}|m] - 2\mathrm{Cov}(\hat{W}, \hat{Z}|m), \end{aligned} \tag{5.4}$$

where we can see that the $\tilde{Z}$'s don't completely cancel.

Recall from (4.14) that the variance of $\hat{W}$ is

$$\mathrm{Var}[\hat{W}|m] = \frac{1}{m^2}\sum_{i=1}^{m} E[W_i^2|m] - \frac{1}{m^2}\sum_{i=1}^{m}\big(E[W_i|m]\big)^2 + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j\neq i}\mathrm{Cov}(W_i, W_j|m).$$

Also recall the sample variance of W_i from (4.15),

$$\frac{1}{m^2}\sum_{i=1}^{m} E\Big[(W_i - \hat{W})^2\Big|m\Big] = \frac{m-1}{m^3}\sum_{i=1}^{m} E[W_i^2|m] - \frac{1}{m^3}\sum_{i=1}^{m}\sum_{j\neq i} E[W_i W_j|m],$$

and we can see that

$$\begin{aligned}
&\mathrm{Var}[\hat{W}|m] - \frac{1}{m^2}\sum_{i=1}^{m} E\Big[(W_i - \hat{W})^2\Big|m\Big] \\
&\quad = \frac{1}{m^2}\sum_{i=1}^{m} E[W_i^2|m] - \frac{1}{m^2}\sum_{i=1}^{m}\big(E[W_i|m]\big)^2 + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j\neq i}\mathrm{Cov}(W_i, W_j|m) \\
&\qquad - \frac{m-1}{m^3}\sum_{i=1}^{m} E[W_i^2|m] + \frac{1}{m^3}\sum_{i=1}^{m}\sum_{j\neq i} E[W_i W_j|m] \\
&\quad = \frac{1}{m^3}\sum_{i=1}^{m} E[W_i^2|m] + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j\neq i}\mathrm{Cov}(W_i, W_j|m) \\
&\qquad + \frac{1}{m^3}\sum_{i=1}^{m}\sum_{j\neq i}\Big(E[W_i W_j|m] - \big(E[W_i|m]\big)^2\Big) - \frac{1}{m^3}\sum_{i=1}^{m}\big(E[W_i|m]\big)^2 \\
&\quad = \frac{1}{m^3}\sum_{i=1}^{m}\mathrm{Var}[W_i|m] + \frac{m+1}{m^3}\sum_{i=1}^{m}\sum_{j\neq i}\mathrm{Cov}(W_i, W_j|m). \qquad (5.5)
\end{aligned}$$

Returning to (5.4) and substituting (5.5) we obtain

$$\begin{aligned}
E[\mathrm{Var}_{\mathrm{Boot}}|m] - \mathrm{Var}[\hat{Y}|m] = &-\frac{1}{m^3}\sum_{i=1}^{m}\mathrm{Var}[W_i|m] - \frac{m+1}{m^3}\sum_{i=1}^{m}\sum_{j\neq i}\mathrm{Cov}(W_i, W_j|m) \\
&+ \frac{2}{m^2}\sum_{i=1}^{m} E\Big[(W_i - \hat{W})(Z_i - \hat{Z})\Big|m\Big] - \frac{1}{m}\tilde{Z} - 2\mathrm{Cov}(\hat{W}, \hat{Z}|m).
\end{aligned}$$

From (4.5),

$$\tilde{Z} = \frac{1}{m^2}\sum_{i=1}^{m} E[Z_i^2|m] \leq \frac{1}{m},$$

since $|Z_i| \leq 1$.

Now we can use (4.18), (4.21), (4.17), and (4.22) to bound each of the terms in the bias,

$$\begin{aligned}
|E[\mathrm{Var}_{\mathrm{Boot}}|m] - \mathrm{Var}[\hat{Y}|m]| &\leq \frac{1}{m^3}\sum_{i=1}^{m}\delta_k^2 + \frac{m+1}{m^3}\sum_{i=1}^{m}\sum_{j\neq i}\delta_k^2 + \frac{2}{m}\delta_k + \frac{1}{m}\Big(\frac{1}{m}\Big) + 2\frac{\delta_k}{\sqrt{m}} \\
&= \delta_k^2 + \frac{2\delta_k}{m} + \frac{1}{m^2} + \frac{2\delta_k}{\sqrt{m}},
\end{aligned}$$

completing the proof of the theorem. □

We would expect the bound to be similar to the jackknife one, since the bootstrap estimate is simple the jackkknife estimate scaled by $\frac{m-1}{m}$. The main difference here is that as we let context length $k \to \infty$, $\delta_k \to 0$, and we are left with $\frac{1}{m^2}$. This is because as we saw in the previous section, the bootstrap estimate is biased. Therefore, we would not expect the bias to disappear completely. The rest of the terms remain the same, except we now have $\frac{2\delta_k}{m}$ instead of $\frac{2\delta_k}{m-1}$, due to the scaling of the estimate. Just as with the jackknife, as $m \to \infty$ the δ_k^2 corresponding to the variance of and covariance between W_i's remains, again because we are approximating the true state $\mathbf{s}$ with an aggregated state $\mathbf{w}$.

Bibliography

[1] M. Asadi, R. P. Torghabeh, and N. P. Santhanam, "Stationary and transition probabilities in slow mixing, long memory markov processes," *IEEE Transactions on Information Theory*, vol. 60, no. 9, pp. 5682–5701, 2014.

[2] B. Efron, *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[3] B. Efron and C. Stein, "The jackknife estimate of variance," *The Annals of Statistics*, pp. 586–596, 1981.

[4] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 01 1979. [Online]. Available: https://doi.org/10.1214/aos/1176344552

[5] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, 1974. [Online]. Available: http://www.jstor.org/stable/2984809

[6] E. Carlstein, "The use of subseries values for estimating the variance of a general statistic from a stationary sequence," *Ann. Statist.*, vol. 14, no. 3, pp. 1171–1179, 09 1986. [Online]. Available: https://doi.org/10.1214/aos/1176350057

[7] H. R. Kunsch, "The jackknife and the bootstrap for general stationary observations," *The Annals of Statistics*, pp. 1217–1241, 1989.

[8] R. Y. Liu and K. Singh, "Moving blocks jackknife and bootstrap capture weak dependence," *Exploring the limits of bootstrap*, vol. 225, p. 248, 1992.

[9] D. N. Politis and J. P. Romano, "A general resampling scheme for triangular arrays of α-mixing random variables with application to the problem of spectral density estimation," *The Annals of Statistics*, pp. 1985–2007, 1992.

[10] ——, "A nonparametric resampling procedure for multivariate confidence regions in time series analysis," in *Computing Science and Statistics*. Springer, 1992, pp. 98–103.

[11] D. N. Politis, J. P. Romano, and T.-L. Lai, "Bootstrap confidence bands for spectra and cross-spectra," *IEEE Transactions on Signal Processing*, vol. 40, no. 5, pp. 1206–1215, 1992.

[12] D. N. Politis and J. P. Romano, "The stationary bootstrap," *Journal of the American Statistical association*, vol. 89, no. 428, pp. 1303–1313, 1994.

[13] J. L. Horowitz, "Bootstrap methods for markov processes," *Econometrica*, vol. 71, no. 4, pp. 1049–1082, 2003.

[14] E. Paparoditis and D. N. Politis, "The local bootstrap for markov processes," *Journal of Statistical Planning and Inference*, vol. 108, no. 1, pp. 301–328, 2002.

[15] B. Hansen, "Non-parametric dependent data bootstrap for conditional moment models," *University of Wisconsin–Madison working paper*, 1999.

[16] P. Bühlmann, "Sieve bootstrap with variable-length markov chains for stationary categorical time series," *Journal of the American Statistical Association*, 2011.

[17] I. Csiszár and Z. Talata, "Consistent estimation of the basic neighborhood of markov random fields," *The Annals of Statistics*, pp. 123–145, 2006.

[18] K. Oshiro, C. Wu, and N. P. Santhanam, "Jackknife estimation for markov processes with no mixing constraints," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 3020–3024.